REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1999	Professional Paper

4. TITLE AND SUBTITLE
Advanced Data Fusion for Wartime Event Correlation and Prediction

5. FUNDING NUMBERS
In-house

6. AUTHOR(S)
M. G. Ceruti, Ph.D., and S. J. McCarthy, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Space and Naval Warfare Systems Center
San Diego, CA 92152–5001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Space and Naval Warfare Systems Center
San Diego, CA 92152-5001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENTR
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This paper summarizes an ongoing data-fusion project the purpose of which is to identify patterns from object-oriented databases derived from message traffic generated during U.S. Marine Corps exercises. These patterns will be used to predict attacks during wartime and other periods of tension or conflict. The paper describes a concept of operations, a literature survey of relevant data-mining and classifer algorithms, plans for system development and directions for future research.

19991203 079

Published in Proceedings of the 23rd Annual IEEE Computer Software and Applications Conference, COMPSAC 99, October 1999.

14. SUBJECT TERMS
Mission Area: Command and Control
Bayesian networks; classifiers; data fusion; data mining algorithms; military exercise; object-oriented database; readiness; Marine Corps

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	Same as Report

DTIC QUALITY INSPECTED 4

Fast Abstract for the Twenty-Third Annual International IEEE Computer Software and Applications Conference, COMPSAC'99, October 27-29, Phoenix, AZ, USA

Advanced Data Fusion for Wartime Event Correlation and Prediction

Marion G. Ceruti, Ph.D., Member,
IEEE Computer Society
Space and Naval Warfare Systems Center, D4121
53560 Hull Street
San Diego, CA 92152-5001, USA
+1 619 553 4068
ceruti@spawar.navy.mil

S. J. McCarthy, Ph.D.
Space and Naval Warfare Systems Center, D432
53560 Hull Street
San Diego, CA 92152-5001, USA
+1 619 553 1520
mccarthy@spawar.navy.mil

Abstract

This paper summarizes an ongoing data-fusion project the purpose of which is to identify patterns from object-oriented databases derived from message traffic generated during U.S. Marine Corps exercises. These patterns will be used to predict attacks during wartime and other periods of tension or conflict. The paper describes a concept of operations, a literature survey of relevant data-mining and classifer algorithms, plans for system development and directions for future research.

Keywords: Bayesian networks, classifiers, data fusion, data mining algorithms, military exercise, object-oriented database, readiness, Marine Corps

1. Introduction

The ability to predict attacks and other hostile events during times of conflict is very desirable to military commanders from the standpoint of readiness. The more advanced notice and the more widespread the notification, the better able all echelons are to respond to threats efficiently and with the correct combination of forces.

The literature is replete with recent research results on data mining and data classification. (See, for example, [2, 3, 4, and 7].) Data mining, data classification and data correlation are among the many techniques used to achieve data fusion. As these techniques mature, better tools become available to model and correlate data from complex operational scenarios.

2. Concept of operation

The concept of operations is to use data-mining and data-classification algorithms to detect patterns associated with attacks (e.g. to identify factors that indicate an imminent attack in the near future) and to correlate them with current events with a view toward supplying military commanders with a prediction of the next attack and a confidence level that pertains to that prediction. A considerable amount of data associated with events that have preceded know attacks is required to model attacks, to search for common features, and to find these patterns in new data.

Success in this effort depends on a characterization of the circumstances that translate to well-defined observables that preceded past attacks. The more detailed the available knowledge, the greater the probability that data instantiating critical variables will be collected. It is expected that such detailed data for all variables will not be available prior to future attacks and that all available data may not be useful in predicting attacks (noise). Thus, the task involves identification of algorithms that can operate on incomplete data; detection of pre-attack features in clutter, and pattern recognition. This project can be successful because modern methods of statistical pattern recognition are sufficiently computationally oriented to use a larger dimensional space and because they are less sensitive to noise.

3. Approach

The approach will include an examination of Marine battlefield intelligence requirements, from which a list of several specific requirements for data fusion with an audit trail will be generated. Hostile events will be characterized with respect to as many relevant variables as are deemed necessary to predict future attacks. Message-traffic databases will be analyzed for the occurrence of telltale signs of pending attacks. A goal is to generate an event prediction (in terms of a probability) with a confidence value associated with it. Therefore, it is necessary to know which combinations of events and observations will have a higher probability of indicating a future attack. A baseline can be modeled from normal operational scenarios.

The attack alarm-generation process and the reduction of false positives can be accomplished using constraints from models of known attacks. One approach is to explore in the generation of a knowledge base based on Bayesian networks.

The generation of the appropriate features that can serve to flag immanent attacks is the least scientific and most challenging part of the process. A literature search will be conducted to determine whether the U.S. Navy or the U.S. Army has made any progress in this area that the Marine Corps can use.

A literature search for data-mining algorithms also is in progress with emphasis on algorithms designed to operate on sparse data or data exceptions. These data-mining algorithms will be used to identify complex patterns in the data that correlate well to hostile events. Criteria for sufficient correlation and confidence levels in data associations will be developed. *Correlation strength*, one metric that could be used, is the ratio of the joint probability to the individual probability of observing a pattern [2].

The Space and Naval Warfare Systems Center has access to SRI's classifier algorithms. For example, the Tree-Augmented Naive Bayes (TAN) is a classifier algorithm based on Bayesian networks developed at SRI with the advantages of robustness and polynomial computational complexity [3 and 4]. Bayesian networks are a suitable technology for the following reasons:

- First, one need not provide all joint probability probability values to specify a probability distribution for collections of independent variables [1].
- Second, one could mix modeling, e.g. explicit knowledge engineering for knowledge elicited from experts, with statistical data induction and adaptivity. This will require fewer data values to induce better quality models.
- Third, one could use these models to compute the value of information. For example, having seen signs "A" and "B" of an imminent attack, what is the best information to collect next to confirm that hypothesis?
- Fourth, one could characterize explicitly the kinds of attacks. For example, given an attack of type "D," what are the most likely signals? These signals could be collected regularly to fill the database used as input into TAN.

TAN makes some tradeoffs between accuracy and computation. It approximates a probability distribution using some constraints on the complexity of the representation; however, it is extremely fast (low polynomial), efficient (one pass over the data), and robust (low order statistics).

TAN accepts data sets as input and induces Bayesian networks as output. Specifically, TAN is intended to be used as a classification algorithm, which means that the input would be a file with tuples of the form $\{x_1, x_2, x_3, ..., x_n, c\}$ where the x_i, are values that variable X_i takes and c is the value that a class (C) variable can take. To set the range of each variable, TAN needs an auxiliary file that includes a description of each variable, including the range of values representing the degree of intensity.

TAN's output is a Bayesian network encoding of $P(C, X_n,...,X_1)$ in an efficient manner. To use TAN as a classifier, one simply computes $P(C|x'_n,....,x'_1)$. Given a new vector $X'_n,...,X'_1$ and having a probability distribution over c, one can select the event with highest probability as the one to classify. To compute confidence on this, the bootstrap method can be used [5].

In addition to TAN, SRI has a more general algorithms for inducing Bayesian networks that do not make the compromises that TAN does. These algorithms try to fit the best distribution possible with no constraints. The disadvantage is that the computation of these models is slower; however, this may be acceptable and desirable in some cases. Algorithms can be implemented with the same data and the results compared.

4. Software implementation

A user-friendly interface will be designed on top of the algorithms to facilitate the selection of the best algorithm to use in a given situation, and to provide automated input of selected data sets to the algorithm of choice. TAN will be used as a base classifier and also as a method to fuse the output of other data-mining and classification algorithms.

Most of the data sets will come from the Integrated Marine Multi-agent Command and Control System (IMMACCS) Database [6].

Acknowledgments

The Space and Naval Warfare Systems Center, San Diego's Science and Technology Initiative provided funding for this project. This work was produced by U.S. government employees as part of their official duties and is not subject to copyright. It is approved for public release with an unlimited distribution.

References

[1] E. Charniak, "Bayesian Networks without Tears," *AI Magazine*, pp. 50-63, Winter 1991.

[2] C. Clifton and R. Steinheiser, "Data Mining on Text," *Proceedings of the 22nd Annual IEEE International 0Computer Software and Applications Conference, COMPSAC'99*, pp. 630-635, Aug. 1998.

[3] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Machine Learning*, vol. 29, no. 2/3, pp. 131-163, Nov./Dec. 1997.

[4] N. Friedman, M. Goldszmidt and T. J. Lee, "Bayesian Network Classification with Continuous Attributes: Getting the Best of Both Discretization and Parametric Fitting," *Proceedings of the International Conference on Machine Learning '98*, ITAD-1632-MS-98-043, 1998.

[5] N. Friedman, M. Goldszmidt and A. Wyner, "On the Application of the Bootstrap for Computing Confidence Measures on Features of Induced Bayesian Networks," *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, 1999.

[6] R. Leighton and J. Pohl, *The IIMACCS Object Model and Database*, IOM Version 1.5, Cal Poly San Luis Obispo, OBDATA00, Nov. 1998

[7] B. M. Thuraisingham, *Data Mining: Technologies, Techniques, Tools and Trends*, CRC Press, Boca Raton, FL., 1999.